RESEARCH IN PROGRESS

Criterion Calibration

How Organizational Policy Shapes Employees' Phishing Decision Thresholds

Yuxiao (Rain) Luo Loyola University Chicago
Adel Yazdanmehr Baruch College, CUNY
Jingguo Wang The University of Texas at Arlington

Dewald Roode Workshop on Information Systems Security Research
University of Arizona, Tucson, October 9–10, 2026

When the phishing test makes the news

“GoDaddy apologizes for ‘insensitive’ phishing email offering bonuses to employees”

CBS News, December 2020

A $650 holiday bonus email turned out to be a security test. About 500 employees clicked and filled in their details.

“West Midlands Trains sends staff phishing email promising bonus”

CNBC, May 2021

Around 2,500 rail staff were thanked for working through the pandemic and promised a payment. A rail union called the test “crass.”

“Embedded training … can make employees even more susceptible to phishing.”

Lain, Kostiainen & Čapkun, IEEE S&P 2022

A 15-month field experiment with more than 14,000 employees of one company.

Organizations are pushing vigilance hard, and employees notice. We know a fair amount about what this does to clicks. We know much less about what it does to legitimate email.

CBS News (Dec 2020); CNBC and The Telegraph (May 2021); Lain, Kostiainen & Čapkun (2022)

Suspicion has a cost that rarely gets counted

About two-thirds

of the messages users report for investigation are not dangerous

Proofpoint, 2021

34.8% vs. 26.1%

benign vs. malicious among 50,000 reports at a bank with about 60,000 employees, over 16 months

Doing et al., EuroUSEC 2024

After training

reports of benign email rose during e-learning and after simulation waves at the same bank

Doing et al., EuroUSEC 2024

What a false alarm costs

- Legitimate email gets delayed or never used
- Coordination slows and the sender is left waiting
- Benign reports add work for IT and security teams
- Past some point, extra suspicion moves error from missed attacks to disrupted work

The managerial question is how to configure the policy that drives suspicion.

Proofpoint, "How Good Are Users at Reporting Phishing Emails?" (2021); Doing, Bárbaro, van der Roest, van Gelder, Zhauniarovich & Parkin (2024)

Phishing research has focused on individual ability

The question the literature asks

"Why are some employees better at telling phishing from legitimate email?"

- Individual differences (Vishwanath et al., 2011; Harrison et al., 2016)
- Message features and influence techniques (Wright et al., 2014)
- Training design (Kumaraguru et al., 2009; Jensen et al., 2017)
- Cognitive processing (Wang et al., 2017; Luo et al., 2025)

Ability matters, but the same employees can show different error patterns under different **organization policies**.

Organization A

Same employees, different policy

A severe penalty for a missed phish. Only misses get counted, so people report whenever they are unsure.

Result: more false alarms (Type I errors).

Organization B

Same employees, different policy

Constant awareness campaigns. Fatigue and reactance pull attention away from the inbox.

Result: more missed attacks (Type II errors).

Open question: which policy configuration produces which error profile, and where is total error cost lowest?

Only the missed phish gets counted

Type II error: a missed phish

Traceable and punishable

- The email is opened or acted on, then logged, attributed and scored
- Security programs are built to count Type II error
- **Consequence:** compromise

Visible to the organization

Type I error: a false alarm

Seen only by the sender

- A legitimate email is reported, put off, or ignored
- Work input goes unused, decisions wait, the sender is left waiting
- **Consequence:** lost productivity

Rarely recorded against anyone

We observe both errors directly.

Our model's Type I and Type II error rates are measured directly from the decision log.

Which error rate a policy configuration increases depends on the mechanism it activates.

Doing et al. (2024); D'Arcy, Hovav & Galletta (2009); Vance et al. (2018)

Policy configuration: the levers an organization controls

Sanction severity

Formal penalty for a missed phishing email (high / low)

- How costly a miss is made for the individual: attribution, a recorded penalty, evaluation
- Draws on deterrence and sanctions research in behavioral security
- Manipulated independently of awareness

Threat-awareness intensity

Frequency and vividness of awareness campaigns (high / low)

- How often and how vividly the threat is kept in front of employees: campaigns, warnings, tests
- Draws on fear-appeal and security-training research
- Manipulated independently of sanctions

2 x 2 Experimental Design

Organizations fall into four quadrants, and we treat the interaction between the two dimensions as a hypothesis

We are still grounding these dimensions in the literature, which will inform their final specification

The base rate is not a lever

Roughly 1–5% of a real inbox is phishing, and the firm does not choose that.

We measure it as a prior belief and prime it only as a robustness check.

D'Arcy, Hovav & Galletta (2009); Johnston & Warkentin (2010); Cram, D'Arcy & Proudfoot (2019); Kumaraguru et al. (2009)

Mechanisms that pull against each other

Policy does not change behavior directly. It works through mechanisms that move the error rates in opposite directions.

Asymmetric blame expectation

Only misses get counted

- A missed phishing email can be traced and punished
- An ignored legitimate email is seen only by its sender
- Employees learn that only misses are noticed

Effect: more reporting, and more false alarms (Type I)

Security fatigue and reactance

Awareness pushed too far

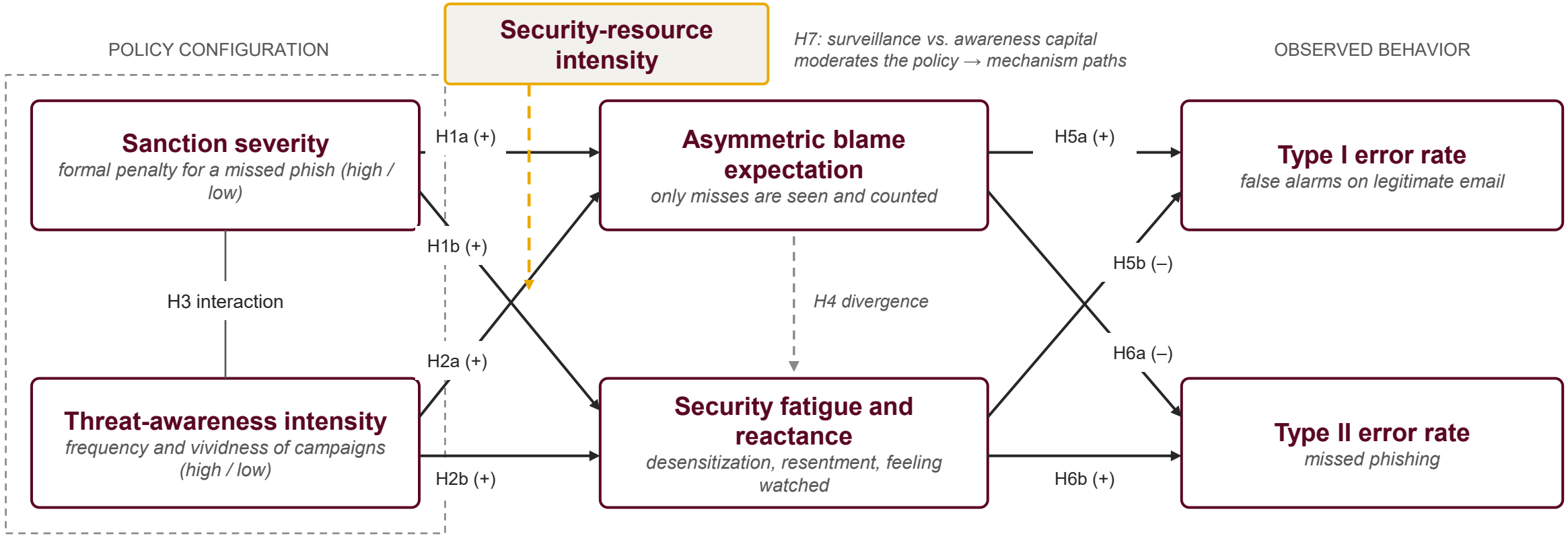
- Desensitization to repeated warnings and tests
- Resentment at being tested and a sense of being watched
- Attention drifts away from the inbox

Effect: more missed phishing (Type II)

Because the mechanisms work against each other, no configuration is clearly best. Total error cost should be lowest somewhere between the extremes.

D'Arcy, Hovav & Galletta (2009); Cram, D'Arcy & Proudfoot (2019); Vance et al. (2018); Reeves et al. (2021); Lain et al. (2022)

Research model



Environment, not a lever: the phishing base rate (about 1–5% of an inbox) is measured as a prior belief and primed only as a robustness check.

Post hoc diagnostic: criterion miscalibration, Δc = c(observed) − c*. It is computed from the same decision log as the error rates, so we report it as a derived diagnostic of where the inflection point falls across quadrants rather than as a path in the model.

Hypotheses: from policy to mechanisms

H1a **Sanction severity → asymmetric blame expectation (+)**
A formal penalty for a missed phish makes the miss the error that counts.

H1b **Sanction severity → security fatigue and reactance (+)**
Penalties also breed resentment at being tested and watched.

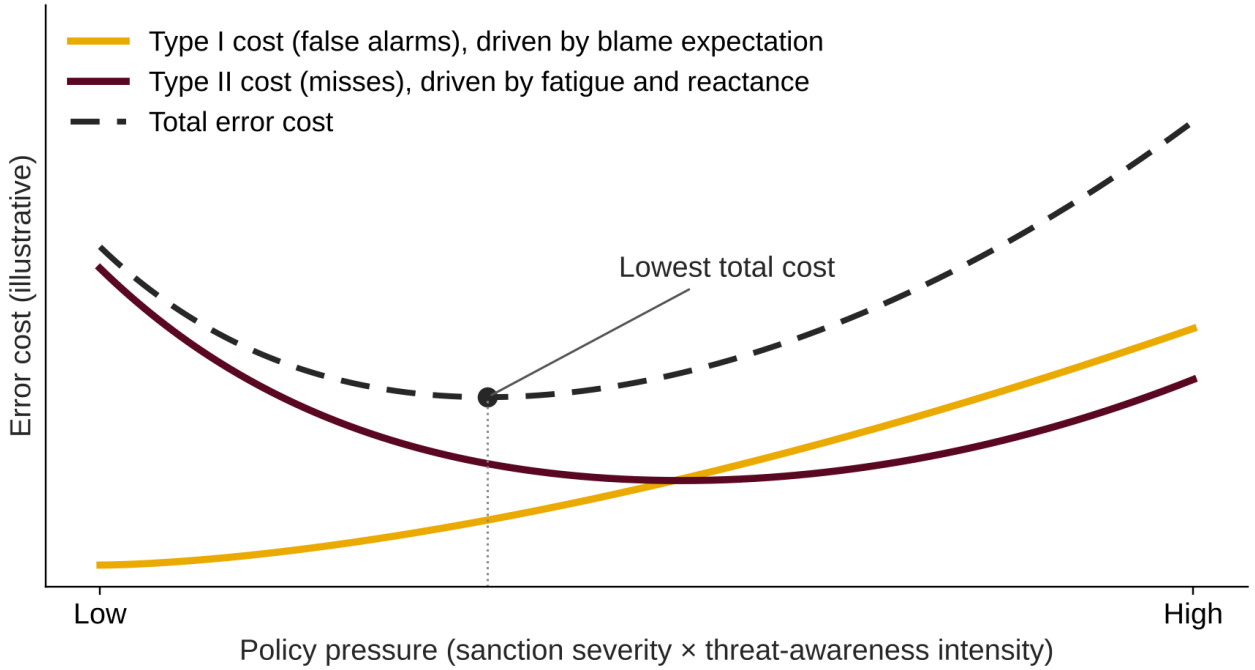
H2a **Threat-awareness intensity → asymmetric blame expectation (+)**
Frequent, vivid campaigns keep the missed phish in mind as the error that matters.

H2b **Threat-awareness intensity → security fatigue and reactance (+)**
Beyond some level, campaigns desensitize people and pull attention from the inbox.

H3 **Configural interaction.** The dimensions are orthogonal by design, and their joint effect on the mechanisms differs across the four quadrants.

H4 **Divergence.** Blame expectation and fatigue/reactance do not move together. They reflect a tension, not one common pressure factor.

Hypotheses: from mechanisms to error rates



Illustrative cost curves implied by the mechanisms.

The study estimates where the minimum actually falls across the four quadrants.

H5 Blame expectation

- **H5a** raises the Type I error rate (false alarms)
- **H5b** lowers the Type II error rate (misses)

H6 Fatigue and reactance

- **H6a** lowers the Type I error rate
- **H6b** raises the Type II error rate

H7 Security-resource intensity moderates the policy → mechanism paths. How the firm splits its budget between surveillance and awareness changes how a configuration pays off.

Study design: a policy-configuration experiment

Sanction severity \ Threat-awareness intensity	low	high
low	**Minimal** Low penalty for a missed phish; infrequent, low-key awareness. **Reference configuration**	**Awareness-led** Frequent, vivid campaigns with no formal penalty. **Both mechanisms rise (H2a, H2b); fatigue expected to dominate**
high	**Deterrence-led** Formal penalty for a missed phish; low-key awareness. **Both mechanisms rise (H1a, H1b); blame expectation expected to dominate**	**Dual-intensity** High penalty and intense awareness. **Configural (H3): where the minimum falls is the empirical question**

Design: 2 × 2 between-subjects

Moderator (H7): security-resource intensity, surveillance vs. awareness capital

Outcomes: Type I and Type II error rates by quadrant, both observed directly

Target sample: 400 complete cases, 100 per cell, carried over from the original design

The base rate stays at its environmental value. It is primed high, low, or not at all as a robustness check, not as a factor.

Experiment Procedure: simulated work session

Step 1
Common training
The same brief phishing training for everyone: definition, common cues, how to report. This evens out declarative knowledge.

Step 2
Baseline block
A short classification block gives baseline false-alarm and miss rates, used as covariates.

Step 3
Policy induction
Random assignment to one of four configurations: sanction severity (high/low) × threat-awareness intensity (high/low).

Step 4
Main task
A simple numerical task whose missing values arrive by email, mixed with phishing. Fewer emails and no pop-ups.

Step 5
Post-task
Mechanism scales, manipulation checks, base-rate prior belief, demand probe, and controls.

Inside the main task
- Each email can be opened and used, reported as phishing, or ignored
- Every legitimate email holds a value the task needs, so a false alarm has a real cost
- Opening a phish is logged silently, with no feedback that could shape later choices

How the outcomes are scored
- Both error rates come from the decision log
- Type I: legitimate emails reported
- Type II: phishing emails opened
- Total error cost: the cost-weighted sum of the two rates

Manipulating each policy dimension

Sanction severity

Formal penalty attached to a missed phishing email

HIGH

A missed phish is attributed to the individual and carries a stated penalty in the evaluation.

LOW

A missed phish is noted without individual penalty; evaluation is aggregate.

Threat-awareness intensity

Frequency and vividness of awareness campaigns

HIGH

Frequent, vivid threat messages and tests before and during the session.

LOW

A single, low-key awareness notice.

Inductions are matched on

- Length
- Visual prominence
- Specificity
- No new diagnostic cues

A pilot will pretest realism, comprehension and strength.

Moderator: security-resource intensity. How the organization splits its security budget between surveillance and awareness capital. We present it as the type of firm participants work for, so we can test whether the same configuration pays off differently across organizations (H7).

Measures and analysis plan follow the causal chain

Measures

Mechanisms	Asymmetric blame expectation; security fatigue and reactance. Parallel scales adapted from sanction-severity, fatigue and reactance measures
Error rates	Type I: legitimate emails reported. Type II: click on phishing emails link. Both from the decision log
Total error cost	Cost-weighted sum of the two rates, with sensitivity analyses over plausible weights
Moderator	Security-resource intensity: surveillance vs. awareness capital, framed as firm type
Base rate	Measured as a prior belief (about 1–5% of an inbox); primed high, low or not at all for robustness

Analysis, in causal order

H1–H3	Effects of the policy dimensions on each mechanism, including the configural interaction
H4	Divergence: no single common factor explains both mechanisms; downstream effects have opposite signs
H5, H6	Structural model with bootstrapped specific indirect effects: configuration → mechanism → error rates
H7	Moderation of the policy → mechanism paths by security-resource intensity, probed with Johnson–Neyman intervals
Inflection point	Total error cost by quadrant, to find the minimum and which configuration suits which organization

What we dropped from the model, and why

The decision criterion

A response threshold is estimated from the same decision log as the error rates, so it cannot come before them. The model now runs straight from policy configuration to the observed error rates.

Fear and frequency overestimation

We removed phishing fear, overestimation of phishing frequency, and cumulative security exposure. The path from exposure to fear was too direct to carry much theory.

The base rate as a lever

The base rate (about 1–5% of an inbox) is set by the environment. We measure it as a prior belief and prime it for robustness. The moderator slot now goes to how the firm spends its security budget.

Boundary conditions

Participants are students in a simulated work session and are less embedded in organizational email routines than employees. Baseline error rates absorb some of that difference, and we state generalization as a boundary. We are still grounding sanction severity and threat-awareness intensity in the literature, and that work will decide the final pair of dimensions.

Contributions and implications for practice

A policy-configuration typology

Anti-phishing policy described by sanction severity and threat-awareness intensity, orthogonal dimensions whose interaction we test directly.

Divergent mechanisms

Asymmetric blame expectation pushes false alarms up, while security fatigue and reactance let misses back in.

An interior optimum

Because the mechanisms diverge, no quadrant is clearly best. The study locates the inflection point and which configuration suits which organization.

A moderator the firm controls

Security-resource intensity, surveillance vs. awareness capital, shapes how a configuration pays off. It replaces the base rate, which the firm never chooses.

FOR PRACTICE

Managing phishing means configuring policy as well as training people.

- The useful question is which configuration fits which kind of organization, more than whether three simulations a year beat five
- Start treating the ignored legitimate email as an error worth recording

Prior phishing research explains individual ability and training. Our explanandum is organizational: the policy configuration, the mechanisms it sets off, and the point where total error cost is lowest.

RESEARCH IN PROGRESS

We would value your feedback on

- Grounding sanction severity and threat-awareness intensity in the literature
- Measuring asymmetric blame expectation and security fatigue and reactance
- Operationalizing security-resource intensity as the moderator
- The simplified numerical task and smaller email stream
- Treating the base rate as a primed robustness check rather than a factor

Thank you

Yuxiao (Rain) Luo
Loyola University Chicago

Adel Yazdanmehr
Baruch College, CUNY

Jingguo Wang
UT Arlington

The base rate is part of the environment

- Real inboxes carry roughly 1–5% phishing. The organization does not choose this.
- So it is not a factor in the model. We measure each participant's prior belief about how common phishing is.
- Robustness check: prime the base rate high, low, or not at all, and confirm that the policy → mechanism → error-rate results hold.
- The moderator is instead a choice the firm actually makes: where it spends its security budget, surveillance or awareness capital (H7).

Why it matters for the task

The task contains more phishing than a real inbox so that both error types can occur in one session. The prior-belief measure and the priming check tell us whether that matters.

Error rates and total error cost

Type I error rate	Share of legitimate emails reported as phishing (or otherwise not used). Each legitimate email holds a value the task needs, so a false alarm has a real productivity cost.
Type II error rate	Share of phishing emails opened or acted on. Opening a phish is logged silently, with no feedback that could shape later decisions.
Total error cost	Cost-weighted sum of the two rates. The primary analysis uses the task's stated costs; sensitivity analyses vary the weight of a false alarm relative to a miss.
Response coding	Primary coding is report vs. open/use. A robustness check recodes "ignore" as a cautious classification and confirms the error-rate results do not change.

Both rates come from the decision log of the main task. Baseline rates from the short pre-task block serve as covariates.

How this study differs from prior phishing research

Prior work

- Individual susceptibility: who falls for phishing and why (Vishwanath et al., 2011; Wright et al., 2014)
- Training effectiveness: whether embedded training or mindfulness reduces clicks (Kumaraguru et al., 2009; Jensen et al., 2017)
- Deterrence and sanctions as drivers of compliance (D'Arcy et al., 2009; Cram et al., 2019)
- Side effects of simulations and warnings such as habituation and fatigue (Vance et al., 2018; Lain et al., 2022; Reeves et al., 2021)
- Each stream looks at one lever or one outcome at a time

This paper

- The explanandum moves from the individual to the organization's policy configuration
- Both policy levers are crossed in one design, with their interaction as a hypothesis
- Divergent mechanisms: blame expectation and fatigue/reactance
- Both error rates are outcomes, which allows an interior optimum instead of "more is better"
- A firm-level moderator: where the security budget goes

Incentives, demand effects and manipulation strength

Isn’t sanction severity just a change in the payoff?

The induction changes the stated personal consequence of a missed phishing email (attribution and a penalty in the evaluation). The performance-based incentive stays the same across conditions. What varies is which error is counted, while the reward rule is unchanged.

Could the inductions work as demand cues?

The inductions are matched in length, prominence and specificity and add no diagnostic cues. A pilot pretests realism, comprehension and strength. After the task we measure perceived experimenter expectations and include the demand probe in robustness analyses.

Is this just a generic monitoring or motivation effect?

The dimensions are manipulated independently, so a generic pressure account predicts mechanisms that move together. H4 predicts divergence, with opposite-signed effects on the error rates, which a single motivation factor cannot produce.

Robustness analyses and controls

Robustness

- Base rate primed high, low, or not at all
- Alternative response coding (ignore as a cautious classification)
- Excluding attention-check failures
- Including demand-probe responses
- Sensitivity analyses over plausible cost weights in total error cost

Controls and checks

- Baseline false-alarm and miss rates (covariates)
- Prior phishing victimization
- Previous exposure to phishing simulations at work or school
- Phishing self-efficacy and task self-efficacy
- Demographics
- Manipulation checks: perceived penalty for a miss; perceived awareness intensity
- Intentions for handling suspicious email in future work

Construct boundaries in the revised model

What each construct is

Sanction severity and threat-awareness intensity are properties of the policy environment, manipulated independently. Asymmetric blame expectation and security fatigue and reactance are employee states produced by that environment. Type I and Type II error rates are behavior observed directly in the decision log. Security-resource intensity is a firm-level choice that moderates how policy produces the mechanisms. The base rate is environmental, measured as a prior belief and primed only for robustness.

Why the mechanisms must diverge

If both mechanisms moved the error rates the same way, more policy would always be better or always worse, and the design question would reduce to how much. Divergence is what makes the inflection point, and the question of which configuration fits which organization, answerable.

Why budget allocation is the moderator

A moderator should be something the organization can act on. The base rate is set by the environment, so it does not qualify. Where the security budget goes, surveillance or awareness capital, is a real choice, and it shapes how the same configuration pays off across firm types (H7).